

Active Label Correction

Umaa Rebbapragada
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA USA
Email: urebbapr@jpl.nasa.gov

Carla E. Brodley
Dept. of Computer Science
Tufts University
Medford, MA USA
Email: brodley@cs.tufts.edu

Damien Sulla-Menashe and Mark A. Friedl
Dept. of Earth and Environment
Boston University
Boston, MA USA
Email: {dsm,friedl}@bu.edu

Abstract—Active Label Correction (ALC) is an interactive method that cleans an established training set of mislabeled examples in conjunction with a domain expert. ALC presumes that the expert who conducts this review is either more accurate than the original annotator or has access to additional resources that ensure a high quality label. A high-cost re-review is possible because ALC proceeds iteratively, scoring the full training set but selecting only small batches of examples that are likely mislabeled. The expert reviews each batch and corrects any mislabeled examples, after which the classifier is retrained and the process repeats until the expert terminates it. We compare several instantiations of ALC to fully-automated methods that attempt to discard or correct label noise in a single pass. Our empirical results show that ALC outperforms single-pass methods in terms of selection efficiency and classifier accuracy. We evaluate the best ALC instantiation on our motivating task of detecting mislabeled and poorly formulated sites within a land cover classification training set from the geography domain.

Keywords-supervised learning, data cleaning, label noise, land cover classification

I. INTRODUCTION

In many scientific domains, training data sets are precious repositories of knowledge that are captured at significant expense. Data may come from ground sensors in remote regions of the planet [1], or from high-cost sensors on board earth-orbiting satellites [2]. Once created, these data sets must be maintained to ensure their usefulness over time.

When testing classification models built from training data, a domain expert may suspect labeling errors to be the source of uncertainty in the results. Labeling errors occur when the annotator spends inadequate time labeling an example, is fatigued, or cannot resolve poor data quality or class ambiguity. The latter occurs frequently in our motivating task of using multi-spectral satellite imagery data to create discrete maps of global land cover using the International Geosphere-Biosphere Programme (IGBP) legend [3]. For example, high-latitude tundra of the “Open Shrubland” class is spectrally similar to high-latitude grasses in the “Grassland” class, thus the two classes are naturally ambiguous and often confused. In this domain, an expert can consult auxiliary data sources (e.g., GoogleEarthTM) to make a high confidence decision for a small set of examples.

However, this incurs significant cost if implemented for all instances during training set creation.

This paper presents Active Label Correction (ALC), a solution to cleaning a long-standing, cultivated training dataset that likely contains labeling errors. ALC was built for the scenario where data owners must pursue the correction of labeling errors (because the generation of new training data is not feasible) and are willing to expend resources to guarantee that the expert performing the “re-review” is either more accurate than the original annotator or has access to additional data sources that will yield new insights. This is a realistic scenario for many domains if the re-review focuses on a small set of examples that are likely mislabeled such that the additional time and cost needed is justified. By highlighting the difficult cases for the re-review, ALC finds truly mislabeled instances that when corrected will result in an “improved” training set, where improvement is measured either by the increase in predictive accuracy of the induced classifier, or the reduction of label noise.

To efficiently utilize the expert’s time, we apply ALC iteratively to present the expert with small batches of suspected mislabeled cases at each round. ALC maintains efficiency by sending for review only those examples it considers likely mislabeled. We evaluate several instantiations of ALC on data sets in which we artificially introduce two types of label noise: random and rule-based class noise [4]. We evaluate each ALC implementation’s ability to present truly mislabeled examples to the expert and compare ALC to fully-automated methods for eliminating and correcting mislabeled examples.

Our results show that ALC efficiently presents examples to the expert that are truly mislabeled, and achieves higher classifier accuracy on more noise types and levels compared to the other approaches. We apply ALC to the motivating domain of this research: a landcover classification data set containing known amounts of label noise, and present results that show the utility of an interactive method in helping domain experts clean training data to further their scientific research objectives.

II. RELATED WORK

Fully-automated label noise detection methods have the goal of removing label noise as a pre-processing step prior to classifier training. The majority of existing *single-pass* label noise techniques discard examples suspected of being mislabeled [5], [6], [7], [8], [9], [10]. Only one attempts to correct labels [11] and two adopt a hybrid approach of both discarding and correcting [12].

Most techniques that detect label noise employ either *confidence-based* [12], [11], [5] or *committee-based* [7], [8], [9], [10] example selection. Confidence-based methods flag examples as mislabeled based on a calculated probability or statistical confidence test. Committee-based approaches build ensembles that flag examples as mislabeled if the majority or consensus vote of ensemble members is against the example's existing label. Ensembles may be created via complementary learning algorithms [7], different feature subsets [8], boosting and bagging [9].

Active learning [13] is distinct from label noise detection in its objective, but also uses confidence- and committee-based example selection. Active learning seeks to generate a maximally-informative training set from a pool of *unlabeled* data. Uncertainty sampling [14] and query by committee (QBC) [15], [16], [17], [18] are effectively confidence-based and committee-based methods. Uncertainty sampling methods select unlabeled examples on which the currently trained classifier is most likely to make classification mistakes. Methods differ on how this determination is made. For example, some estimate the posterior probability of class membership [14], while others use heuristics specific to a particular classifier [19], or a hybrid approach [20]. QBC methods select examples for labeling by querying a committee of classifiers trained on the current set of labeled data, and choosing those examples that show maximal disagreement.

The work with the most similar goal is the multiple expert active learning scenario [21], in which there are multiple experts of differing costs and expertise and the goal is to determine who should label each candidate instance. Wallace et al. find that the best strategy is to have the less expensive expert indicate which instances the more expensive (and presumably more accurate) expert should label. In contrast, ALC automatically determines which examples to re-examine rather than relying on a human.

III. ACTIVE LABEL CORRECTION

Figure 1 presents the algorithm for ALC. ALC takes as input the original training set T containing an unknown amount of label noise, the number of examples k to select at each round, and a selection algorithm that determines which examples are most likely mislabeled. At round i , the selection algorithm chooses k previously unexamined examples from $T_i - S_C$ to present to the expert, where S_C is the set of examples previously examined and corrected.

Figure 1. Active Label Correction

```

1: Inputs:  $T$  (training set),  $k$  (batch-size), Select (selection
   method)
2:  $S_C \leftarrow \emptyset$ ;  $T_0 \leftarrow T$ ;  $i \leftarrow 0$ 
3: repeat
4:    $S_i \leftarrow \text{Select}(T_i, S_C, k)$   $\{S_i \cap S_C = \emptyset\}$ 
5:   Expert-Relabels( $S_i$ )
6:    $T_{i+1} \leftarrow \text{Update-Labels}(T_i, S_i)$ 
7:    $S_C \leftarrow S_C \cup S_i$ 
8:    $i \leftarrow i + 1$ 
9: until expert is finished

```

The expert reviews example set S_i and corrects any that are determined to be mislabeled. The training set is then updated to reflect the new labels, the selected examples are added to S_C , and the process repeats.

The termination of ALC is left to the expert's discretion. We expect the expert to make a decision to terminate based on the efficiency of the selection algorithm in presenting truly mislabeled examples for review. Ideally, all or most of the k examples presented for review in the early rounds of ALC will be ones that the expert agrees are mislabeled. As ALC progresses, one would expect the proportion of truly mislabeled examples in the batch to decrease. The expert may terminate the process when the number of clean examples exceeds the number of mislabeled examples in the batch or when the given relabeling budget has been exhausted.

A. Selection Methods

We describe two methods for scoring called *ALC-Mislabeled* and *ALC-Disagreement*. The ALC-Mislabeled score estimates how likely an example x is mislabeled by calculating the difference in probabilities of the existing and predicted label. Assuming that for each x we use a predictive method that outputs a probability distribution P over the set of class labels, we compute the ALC-Mislabeled score as:

$$\text{Score}_{\text{Mislabeled}}(x) = P(l_p) - P(l_e)$$

where l_e is the existing label and l_p is the predicted label (with maximum probability). The larger (more positive) the value of $\text{Score}_{\text{Mislabeled}}(x)$, the more likely it is that x is mislabeled. We sort $\text{Score}_{\text{Mislabeled}}$ in descending order and choose the largest k scores for expert review.

Our second scoring method, ALC-Disagreement, chooses examples for relabeling that are not obviously mislabeled; it selects examples that exhibit a large degree of confusion as to the predicted label and thus can be viewed as selecting "hard" to classify examples. This confusion is reflected by the probability distribution over the class labels – the closer it is to a uniform distribution the more confusion. This is the idea behind many active learning methods that seek to find the samples that would provide the most benefit if

labeled (or in the case of ALC, the most benefit if relabeled). To see how this situation differs from ALC-Mislabeled's motivation, consider an example for which $P(l_p) = 0.9$ and $P(l_e) = 0.1$. This example incurs a high ALC-Mislabeled score, but the classifier is highly confident of its prediction even though it disagrees with the assignment label. Clearly this label should be corrected but it's possible this example is not particularly informative to classification. Based on these observations we calculate ALC-Disagreement as:

$$Score_{Disagreement}(x) = - \sum_i P(l_i) \log P(l_i)$$

This strategy is directly modeled on active learning methods that use the entropy of P to measure the uncertainty of the classifier's prediction [16], [22].

B. Calculating Label Probabilities

We describe six methods for obtaining P , the probability distribution over class labels on each example, four of which are confidence-based and two of which are committee-based. Our preference is to use natural or calibrated [23] classifier probabilities for confidence-based selection methods, and calculate P for each as follows:

- Naive Bayes (NB): Native probabilities.
- Logistic Regression (LOG): Native probabilities.
- SMO: Calibrated probabilities where a logistic function is fit over the SVM outputs [24], [23].
- Decision trees (J48): A probability distribution is derived from class label counts at the node level.

Given the alternatives available for committee-based selection [18], [17], [9], [8], [7], we choose the Query-by-Bagging (BAG) and Query-by-Boosting (BST) method of committee creation [18]. These have performed well in previous evaluations, and have the advantage of being computationally tractable and compatible with any learning algorithm. We implement BAG and BST as follows:

- BAG: A committee of ten decision tree (C4.5) classifiers trained on a random sampling (2/3, with replacement) of the training set.
- BST [25]: A committee of ten sequential decision tree (C4.5) classifiers where each classifier is built with a random sampling (2/3) of the training set and instances are resampled according to a weighting scheme that concentrates on examples that are classified incorrectly by the previous classifier.

For committee-based methods BAG and BST, ALC-Mislabeled and ALC-Disagreement scores are calculated:

$$Score_{Mislabeled} = \hat{P}(l_p) - \hat{P}(l_e)$$

$$Score_{Disagreement}(x) = - \sum_i \hat{P}(l_i) \log \hat{P}(l_i)$$

where \hat{P} is the mean probability distribution derived from the committee.

	Segmentation (S)	Road (R)	Land Cover (L)
RU10	5.84 ± 0.55	10.06 ± 0.91	6.69 ± 0.49
RU20	11.26 ± 0.68	20.02 ± 1.09	13.36 ± 0.58
RU30	17.25 ± 0.78	29.84 ± 1.21	19.90 ± 0.77
RU40	22.81 ± 0.95	39.67 ± 1.30	26.50 ± 0.90

Table I
ACTUAL PERCENTAGES CORRESPONDING TO POTENTIAL NOISE LEVELS RU10, 20, 30 AND 40 FOR RULE-BASED NOISE.

For each of the six methods, we train a *single* classifier using *all* of the training data.

IV. EXPERIMENTS

Our primary experimental goal is to compare ALC-Mislabeled and ALC-Disagreement with an ALC that randomly selects examples (ALC-Rand), and both single-pass discarding (SPD) and correcting (SPC). We compare each method's performance in terms of selection efficiency and classification accuracy on data in which we simulate two types of label noise.

A. SPD, SPC and ALC Implementations

SPC and SPD are fully-automated techniques that attempt to eliminate noise in a single pass of the data by correcting and discarding examples respectively. SPC and SPD are capable of two types of errors: 1) not correcting or discarding a truly mislabeled example and 2) correcting or discarding a truly clean example. Indeed, SPC and SPD can introduce more noise into the training data if too many truly clean examples are discarded or mislabeled.

We implement SPC, SPD and both ALC-Mislabeled and ALC-Disagreement with label probability methods LOG, SMO, NB, J48, BAG and BST. For SPD, we discard examples whose probabilities or committee votes on the existing label are less than the probabilities or votes on the predicted label. For SPC, we simply update the example to have the predicted label, which is the label receiving the highest probability or highest vote.

B. Simulating Label Noise

We simulate label noise on three (labeled) data sets, referred to as Segmentation (S) [26], Road (R), and Land Cover (L) [27]. These are multi-class data sets with 2310, 2056 and 3398 instances and 7, 9 and 11 classes respectively. These are the same data sets used in [7], [4]. For detailed descriptions, we refer the reader to [7].

We introduce two types of label noise into the data: random and rule-based noise. Random noise is introduced by randomly selecting $n\%$ of the training set in proportion to its class distribution. For each instance selected for noise injection, we flip its label to one chosen uniformly among the other classes. We denote random noise levels of 10, 20, 30 and 40% as RA10, RA20, RA30 and RA40. Rule-based noise is introduced with the assistance of rules provided

by a domain expert that reflect natural confusions between classes. We use the rules outlined in [7]. Each instance has an $n\%$ chance of being flipped according to its rule. Thus, the data set potentially has $n\%$ noise, but the actual noise level is less than $n\%$ because there may not be a rule for each pair of classes. Mislabeling to ensure $n\%$ noise in the data can create pathological mislabeling among minority classes in cases where minority classes have mislabeling rules and majority classes do not. For rule-based noise, RU10, RU20, RU30 and RU40 indicate the potential noise levels of 10, 20, 30 and 40%.

We generated 50 versions of each data set at each noise level. Table I reports the mean actual noise levels. Note that actual noise levels for Segmentation and Land Cover are below potential noise levels because not all classes have mislabeling rules.

C. Experimental Method

For ALC, we simulate a domain expert that perfectly corrects and does not re-introduce noise into the data. We select instances for correcting in batches (k) of 20 (chosen arbitrarily). We terminate each experimental run after $\lfloor n/20 \rfloor$ rounds where n is the number of truly mislabeled examples in the training set.

We used WEKA [28] to implement our classifiers. For each of the 50 runs we randomly shuffle the data set and reserve the first 2/3 for training and final 1/3 for testing. We introduced noise into the training sets only; test sets remained noise free.

D. Selection Efficiency

We analyzed the selection efficiency of each label probability method discussed in Section III-B, but omit the results due to space reasons. Our results show that ALC example selection performs better than random in all cases, and that the overall winners are BAG, LOG and SMO.

Figure 2 compares ALC-Mislabeled and ALC-Disagreement in terms of ability to select truly mislabeled examples (precision) for all data sets and noise types, and reports the result using the best-performing label probability method for each ALC method. For both noise models (RA and RU) at all noise levels, as one would expect, ALC-Mislabeled is able to more accurately find mislabeled instances than ALC-Disagreement. This result is consistent with results presented in [29] where label uncertainty (LU) scoring outperformed model uncertainty (MU) scoring for two class tasks in the presence of high quality labeling.

E. Predictive Accuracy Results

In Figure 3 we compare automated label discard and correction methods (SPD and SPC) to ALC-Rand, ALC-Disagreement and ALC-Mislabeled with respect to the predictive accuracy. We also note how each method improves

predictive accuracy over the baseline of doing nothing (mislabeled kept or MK in the graphs). For increased clarity in the figures, we only report the results obtained with the best performing label probability method. All results are generated using the J48 decision tree algorithm, but we also generated these results using an SVM as the final classifier with comparable results.

All noise correction methods, including ALC-Rand perform better than MK at all noise levels with the exception of the 10% noise-level. ALC-Rand, although better than doing nothing, performs more poorly than all other interventions. SPD and SPC perform more poorly than ALC methods at higher levels of noise on the Segmentation and Land Cover datasets. ALC-Mislabeled and ALC-Disagreement have comparable results for all levels of random noise, but ALC-Mislabeled performs slightly better than ALC-Disagreement for rule-based noise on the Segmentation and Land Cover datasets at higher levels of noise. Although this difference in predictive accuracy is not large, it is important to note that these methods will interact with a human expert who is likely to be more engaged in spending effort relabeling if he/she is indeed changing labels. Even though the differences in accuracy are small, the differences in precision (as reported in the last section) are significant enough that one should use ALC-Mislabeled as the selection method.

F. ALC Applied to Land Cover Classification

We apply ALC to a database of known land cover types used to train a global land cover classifier. The source of the data is the System for Terrestrial Ecosystem Parameterization (STEP) database [30], which is a collection of about 2000 polygons drawn over known land cover types produced from data collected by the Moderate Resolution Imaging Spectroradiometer (MODIS).

The labeling scheme is determined by the International Geosphere-Biosphere Programme [30], which is the consensus standard for the Earth science modeling community. The IGBP legend contains seventeen land cover classes. Ideally, the land cover classes should be mutually exclusive and distinct with respect to both space and time. By this criteria, the IGBP scheme is not ideal because it tries to characterize land cover, land use, and surface hydrology, which may overlap in certain areas. The version of STEP used to produce the MODIS Collection 5 Land Cover product was initially labeled at Boston University between 1998-2005 using various information sources such as Landsat TM imagery. Since then new information sources have become available such as high resolution GoogleEarthTM imagery, independently produced maps of agriculture, an extended time series of MODIS-derived vegetation indices, and ecoregion maps [3]. These new sources of information make it an ideal candidate for ALC.

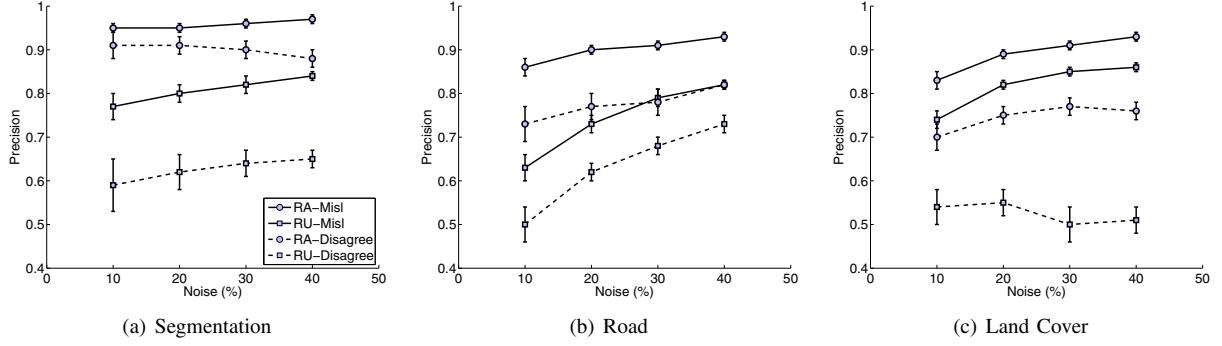


Figure 2. Precision of ALC-Mislabeled and ALC-Disagreement (ability to find true mislabeled examples among examples selected).

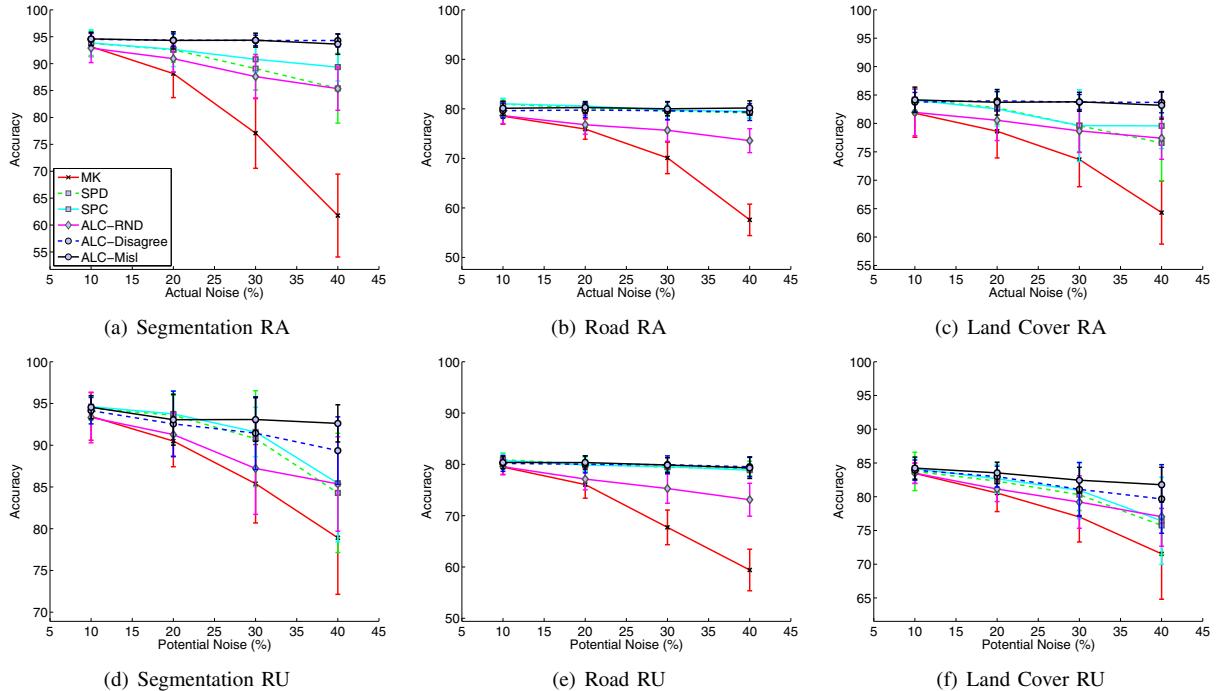


Figure 3. Results on test set accuracy using a decision tree (J48) classifier.

Our experiment ran six rounds of ALC-Mislabeled; at each round we asked the domain expert to examine the ten sites selected by ALC implemented with the SMO label probability method.¹ In total, our expert examined sixty unique sites and provided his personal primary and secondary labeling for each site, the confidence on the composition of the site itself, and some detailed comments.

ALC achieved 56.6% precision using our expert's primary labels (details of this analysis can be found in [31]). Informally, the expert told us that "[ALC] is both useful and important to the maintenance of the database. (1) It identifies sites that are very heterogeneous and mixtures of more than one land cover class. (2) It identifies sites that are mislabeled

or of poor quality. I would like to see this analysis on some of the other STEP labels besides the IGBP class since the IGBP classification has a number of problems with class ambiguity."

V. CONCLUSION

This paper presents a human/machine learning method for cleaning training data of label noise when an alternate, high confidence labeling source is available. Within the ALC approach we experimented with two scoring methods and a variety of label probability methods inspired by past work in active learning and label noise detection. We presented precision and accuracy results on several real-world data sets into which we introduced both random noise and noise in the manner in which it would naturally occur. We compared

¹The SVM used a polynomial kernel (order 2) where the C parameter was set to 1.0.

ALC against a variety of baselines, including random selection of examples and single-pass correcting and discarding. Our results show that involving the expert in a clever way outperforms automated data cleaning methods. Our research was motivated by the task of updating landcover classification data given that new sources of information are available that were not available when the datasets were compiled.

ACKNOWLEDGMENT

The research described in this paper took place while Umaa Rebbapragada was a graduate student at Tufts University. Carla Brodley's and Umaa Rebbapragada's research was supported by NSF grants IIS-0803409 and IIS-0713259.

REFERENCES

- [1] R. Aster, W. McIntosh, and P. Kyle et al., "Real-time data received from Mount Erebus volcano, Antarctica," *Eos*, vol. 85, no. 10, pp. 97–104, 2004.
- [2] L. Mandrake, K. Wagstaff, and D. Gleeson et al., "Onboard detection of naturally occurring sulfur compounds on the surface of a glacier using an svm and the hyperion multispectral instrument," in *IEEE Aerospace Conference*, 2009.
- [3] M. A. Friedl, D. Sulla-Menashe, and B. Tan et al., "MODIS collection 5 global land cover: Algorithm refinements and characterization of new datasets," *Remote Sensing of Environment*, vol. 114, no. 1, pp. 168–182, 2010.
- [4] U. Rebbapragada and C. E. Brodley, "Class noise mitigation through instance weighting," in *ECML*, 2007, pp. 708–715.
- [5] X. Zeng and T. R. Martinez, "A noise filtering method using neural networks," in *Proc. of the Int. Wkshp. of Soft Computing Techniques in Instrumentation, Measurement and Related Applications*, 2003.
- [6] D. Gamberger, N. Lavrač, and C. Grošelj, "Experiments with noise filtering in a medical domain," in *ICML*, 1999, pp. 143–151.
- [7] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *JAIR*, vol. 11, pp. 131–167, 1999.
- [8] S. Venkataraman, D. Metaxas, and D. Fradkin et al., "Distinguishing mislabeled data from correctly labeled data in classifier design," in *ICTAI*, 2004, pp. 668–672.
- [9] S. Verbaeten and A. V. Assche, "Ensemble methods for noise elimination in classification problems," in *Multiple Classifier Systems, 4th International Workshop*, 2003.
- [10] X. Zhu, X. Wu, and S. Chen, "Eliminating class noise in large datasets," in *ICML*, 2003, pp. 920–927.
- [11] X. Zeng and T. Martinez, "An algorithm for correcting mislabeled data," *Intelligent Data Analysis*, vol. 5, no. 1, 2001.
- [12] F. Muhlenbach, S. Lallich, and D. A. Zighed, "Identifying and handling mislabelled instances," *Journal of Intelligent Information Systems*, vol. 22, no. 1, pp. 89–109, 2004.
- [13] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.
- [14] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR*, 1994, pp. 3–12.
- [15] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *COLT*, 1992, pp. 287–294.
- [16] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," *ICML*, pp. 150–157, 1995.
- [17] Y. Freund, H. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, no. 2-3, pp. 133–168, 1997.
- [18] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *ICML*, 1998, pp. 1–9.
- [19] S. Tong and D. Koller, "Support vector machine learning with applications to text classification," *JMLR*, vol. 2, pp. 45–66, 2001.
- [20] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," in *ICML*, 2003, pp. 19–26.
- [21] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Who should label what? Instance allocation in multiple expert active learning," in *SDM*, 2011, pp. 176–187.
- [22] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *ICML*, 2004, pp. 584–591.
- [23] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *ICML*, 2005, pp. 625–632.
- [24] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.
- [25] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, 1996, pp. 148–156.
- [26] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [27] R. S. Defries and J. R. G. Townshend, "NDVI-derived land cover classifications at a global scale," *International Journal of Remote Sensing*, vol. 15, no. 17, pp. 3567–3586, 1994.
- [28] I. H. Witten and E. Frank, *Data mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.
- [29] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *KDD*, 2008, pp. 614–622.
- [30] D. Muchoney, A. Strahler, J. Hodges, and J. LoCastro, "The IGBP DISCover confidence sites and the system for terrestrial ecosystem parameterization: Tools for validating global land cover data," *Photogrammetric Engineering & Remote Sensing*, vol. 65, no. 9, 1999.
- [31] U. Rebbapragada, "Strategic targeting of outliers for expert review," Ph.D. dissertation, Tufts University, 2010.